

统计学习方法

李航 著

清华大学出版社


统计学习方法

李航 著



李航 日本京都大学电气工程系毕业，日本东京大学计算机科学博士。曾任职于日本NEC公司中央研究所，现任微软亚洲研究院高级研究员及主任研究员。北京大学、南开大学、西安交通大学客座教授。研究方向包括信息检索、自然语言处理、统计机器学习及数据挖掘。

清华大学出版社数字出版网站

WQBook  书文局

www.wqbook.com

ISBN 978-7-302-27595-4



9 787302 275954 >

定价：38.00元

统计学习方法

李航 著

清华大学出版社
北京

内 容 简 介

统计学习是计算机及其应用领域的一门重要的学科。本书全面系统地介绍了统计学习的主要方法,特别是监督学习方法,包括感知机、 k 近邻法、朴素贝叶斯法、决策树、逻辑斯谛回归与最大熵模型、支持向量机、提升方法、EM算法、隐马尔可夫模型和条件随机场等。除第1章概论和最后一章总结外,每章介绍一种方法。叙述从具体问题或实例入手,由浅入深,阐明思路,给出必要的数学推导,便于读者掌握统计学习方法的实质,学会运用。为满足读者进一步学习的需要,书中还介绍了一些相关研究,给出了少量习题,列出了主要参考文献。

本书是统计学习及相关课程的教学参考书,适用于高等院校文本数据挖掘、信息检索及自然语言处理等专业的大学生、研究生,也可供从事计算机应用相关专业的研发人员参考。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

统计学习方法 / 李航著. --北京:清华大学出版社, 2012.3

ISBN 978-7-302-27595-4

I. ①统… II. ①李… III. ①机器学习 IV. ①TP181

中国版本图书馆 CIP 数据核字(2011)第 270938 号

责任编辑:薛 慧

封面设计:薛 慧

责任校对:王淑云

责任印制:张雪娇

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座 邮 编:100084

社总机:010-62770175 邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

印 刷 者:三河市君旺印装厂

装 订 者:三河市新茂装订有限公司

经 销:全国新华书店

开 本:165mm×240mm 印 张:15.75 字 数:314 千字

版 次:2012 年 3 月第 1 版 印 次:2012 年 3 月第 1 次印刷

印 数:1~3000

定 价:38.00 元

产品编号:023367-01

献给我的母亲

前 言

计算机与网络已融入到了人们的日常学习、工作和生活之中，成为人们不可或缺的助手和伙伴。计算机与网络的飞速发展完全改变了人们的学习、工作和生活方式。智能化是计算机研究与开发的一个主要目标。近几十年来的实践表明，统计机器学习方法是实现这一目标的最有效手段，尽管它还存在着一定的局限性。

作者一直从事利用统计学习方法对文本数据进行各种智能性处理的研究，包括自然语言处理、信息检索、文本数据挖掘。近 20 年来，这些领域发展之快，应用之广，实在令人惊叹！可以说，统计机器学习是这些领域的核心技术，在这些领域的发展及应用中起着决定性的作用。

作者在日常的研究工作中经常指导学生，并在国内外一些大学及讲习班上多次做过关于统计学习的报告和演讲。在这一过程中，同学们学习热情很高，希望得到指导，这使作者产生了撰写本书的想法。

国内外已出版了多本关于统计机器学习的书籍，比如，Hastie 等人的《统计学习基础》。该书对统计学习的诸多问题有非常精辟的论述，但对初学者来说显得有些深奥。统计学习范围甚广，一两本书很难覆盖所有问题。本书主要是面向将统计学习方法作为工具的科研人员与学生，特别是从事信息检索、自然语言处理、文本数据挖掘及相关领域的研究与开发的科研人员与学生。

本书力求系统而详细地介绍统计学习的方法。在内容选取上，侧重介绍那些最重要、最常用的方法，特别是关于分类与标注问题的方法。对其他问题及方法，如聚类等，计划在今后的写作中再加以介绍。在叙述方式上，每一章讲述一种方法，各章内容相对独立、完整；同时力图用统一框架来论述所有方法，使全书整体不失系统性。读者可以从头到尾通读，也可以选择单个章节细读。对每一方法的讲述力求深入浅出，给出必要的推导证明，提供简单的实例，使初学者易于掌握方法的基本内容，领会方法的本质，并准确地使用方法。对相关的深层理论，则仅予以简述。在每章后面，给出一些习题，介绍一些相关的研究动向和阅读材料，列出参考文献，以满足读者进一步学习的需求。本书第 1 章简要叙述统计学习方法的基本概念，最后一章对统计学习方法进行比较与总结。此外，在附录中简要介绍一些共用的最优化理论与方法。

本书可以作为统计机器学习及相关课程的教学参考书，适用于信息检索及自然语言处理等专业的大学生、研究生。

本书初稿完成后，田飞、王佳磊、武威、陈凯、伍浩铨、曹正、陶宇等人分别审阅了全部或部分章节，提出了许多宝贵意见，对本书质量的提高有很大帮

助。在此向他们表示衷心的感谢。在本书写作和出版过程中，清华大学出版社的责任编辑薛慧给予了很多帮助，在此特向她致谢。

由于作者水平所限，书中难免有错误和不当之处，欢迎专家和读者给予批评指正。来函请发至 ml-book-hangli@hotmail.com。

李 航

2011 年 4 月 23 日

符号表

\mathbf{R}	实数集
\mathbf{R}^n	n 维实数向量空间, n 维欧氏空间
\mathbf{H}	希尔伯特空间
\mathbf{X}	输入空间
\mathbf{Y}	输出空间
$x \in \mathbf{X}$	输入, 实例
$y \in \mathbf{Y}$	输出, 标记
X	输入随机变量
Y	输出随机变量
$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$	训练数据集
N	样本容量
(x_i, y_i)	第 i 个训练数据点
$x = (x^{(1)}, x^{(2)}, \dots, x^{(n)})^T$	输入向量, n 维实数向量
$x_i^{(j)}$	输入向量 x_i 的第 j 分量
$P(X), P(Y)$	概率分布
$P(X, Y)$	联合概率分布
\mathbf{F}	假设空间
$f \in \mathbf{F}$	模型, 特征函数
θ, w	模型参数
$w = (w_1, w_2, \dots, w_n)^T$	权值向量
b	偏置
$J(f)$	模型的复杂度
R_{emp}	经验风险或经验损失
R_{exp}	风险函数或期望损失
L	损失函数, 拉格朗日函数
η	学习率
$\ \cdot\ _1$	L_1 范数
$\ \cdot\ _2, \ \cdot\ $	L_2 范数
$(x \cdot x')$	向量 x 与 x' 的内积

$H(X), H(p)$	熵
$H(Y X)$	条件熵
S	分离超平面
$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$	拉格朗日乘子, 对偶问题变量
α_i	对偶问题的第 i 个变量
$K(x, z)$	核函数
$\text{sign}(x)$	符号函数
$I(x)$	指示函数
$Z(x)$	规范化因子

目 录

第 1 章	统计学习方法概论	1
1.1	统计学习	1
1.2	监督学习	3
1.2.1	基本概念	4
1.2.2	问题的形式化	5
1.3	统计学习三要素	6
1.3.1	模型	6
1.3.2	策略	7
1.3.3	算法	9
1.4	模型评估与模型选择	10
1.4.1	训练误差与测试误差	10
1.4.2	过拟合与模型选择	11
1.5	正则化与交叉验证	13
1.5.1	正则化	13
1.5.2	交叉验证	14
1.6	泛化能力	15
1.6.1	泛化误差	15
1.6.2	泛化误差上界	15
1.7	生成模型与判别模型	17
1.8	分类问题	18
1.9	标注问题	20
1.10	回归问题	21
	本章概要	23
	继续阅读	23
	习题	23
	参考文献	24
第 2 章	感知机	25
2.1	感知机模型	25
2.2	感知机学习策略	26

2.2.1	数据集的线性可分性	26
2.2.2	感知机学习策略	26
2.3	感知机学习算法	28
2.3.1	感知机学习算法的原始形式	28
2.3.2	算法的收敛性	31
2.3.3	感知机学习算法的对偶形式	33
	本章概要	35
	继续阅读	36
	习题	36
	参考文献	36
第 3 章	k 近邻法	37
3.1	k 近邻算法	37
3.2	k 近邻模型	38
3.2.1	模型	38
3.2.2	距离度量	38
3.2.3	k 值的选择	40
3.2.4	分类决策规则	40
3.3	k 近邻法的实现: kd 树	41
3.3.1	构造 kd 树	41
3.3.2	搜索 kd 树	42
	本章概要	44
	继续阅读	45
	习题	45
	参考文献	45
第 4 章	朴素贝叶斯法	47
4.1	朴素贝叶斯法的学习与分类	47
4.1.1	基本方法	47
4.1.2	后验概率最大化的含义	48
4.2	朴素贝叶斯法的参数估计	49
4.2.1	极大似然估计	49
4.2.2	学习与分类算法	50
4.2.3	贝叶斯估计	51

本章概要	52
继续阅读	53
习题	53
参考文献	53
第 5 章 决策树	55
5.1 决策树模型与学习	55
5.1.1 决策树模型	55
5.1.2 决策树与 if-then 规则	56
5.1.3 决策树与条件概率分布	56
5.1.4 决策树学习	56
5.2 特征选择	58
5.2.1 特征选择问题	58
5.2.2 信息增益	60
5.2.3 信息增益比	63
5.3 决策树的生成	63
5.3.1 ID3 算法	63
5.3.2 C4.5 的生成算法	65
5.4 决策树的剪枝	65
5.5 CART 算法	67
5.5.1 CART 生成	68
5.5.2 CART 剪枝	72
本章概要	73
继续阅读	75
习题	75
参考文献	75
第 6 章 逻辑斯谛回归与最大熵模型	77
6.1 逻辑斯谛回归模型	77
6.1.1 逻辑斯谛分布	77
6.1.2 二项逻辑斯谛回归模型	78
6.1.3 模型参数估计	79
6.1.4 多项逻辑斯谛回归	79
6.2 最大熵模型	80
6.2.1 最大熵原理	80
6.2.2 最大熵模型的定义	82

6.2.3	最大熵模型的学习	83
6.2.4	极大似然估计	87
6.3	模型学习的最优化算法	88
6.3.1	改进的迭代尺度法	88
6.3.2	拟牛顿法	91
	本章概要	92
	继续阅读	93
	习题	94
	参考文献	94
第 7 章	支持向量机	95
7.1	线性可分支持向量机与硬间隔最大化	95
7.1.1	线性可分支持向量机	95
7.1.2	函数间隔和几何间隔	97
7.1.3	间隔最大化	99
7.1.4	学习的对偶算法	103
7.2	线性支持向量机与软间隔最大化	108
7.2.1	线性支持向量机	108
7.2.2	学习的对偶算法	110
7.2.3	支持向量	113
7.2.4	合页损失函数	113
7.3	非线性支持向量机与核函数	115
7.3.1	核技巧	115
7.3.2	正定核	118
7.3.3	常用核函数	122
7.3.4	非线性支持向量分类机	123
7.4	序列最小最优化算法	124
7.4.1	两个变量二次规划的求解方法	125
7.4.2	变量的选择方法	128
7.4.3	SMO 算法	130
	本章概要	131
	继续阅读	133
	习题	134
	参考文献	134

第 8 章 提升方法	137
8.1 提升方法 AdaBoost 算法	137
8.1.1 提升方法的基本思路	137
8.1.2 AdaBoost 算法	138
8.1.3 AdaBoost 的例子	140
8.2 AdaBoost 算法的训练误差分析	142
8.3 AdaBoost 算法的解释	143
8.3.1 前向分步算法	144
8.3.2 前向分步算法与 AdaBoost	145
8.4 提升树	146
8.4.1 提升树模型	147
8.4.2 提升树算法	147
8.4.3 梯度提升	151
本章概要	152
继续阅读	153
习题	153
参考文献	153
第 9 章 EM 算法及其推广	155
9.1 EM 算法的引入	155
9.1.1 EM 算法	155
9.1.2 EM 算法的导出	158
9.1.3 EM 算法在非监督学习中的应用	160
9.2 EM 算法的收敛性	160
9.3 EM 算法在高斯混合模型学习中的应用	162
9.3.1 高斯混合模型	162
9.3.2 高斯混合模型参数估计的 EM 算法	163
9.4 EM 算法的推广	166
9.4.1 F 函数的极大-极大算法	166
9.4.2 GEM 算法	168
本章概要	169
继续阅读	170
习题	170
参考文献	170

第 10 章 隐马尔可夫模型	171
10.1 隐马尔可夫模型的基本概念	171
10.1.1 隐马尔可夫模型的定义	171
10.1.2 观测序列的生成过程	174
10.1.3 隐马尔可夫模型的 3 个基本问题	174
10.2 概率计算算法	174
10.2.1 直接计算法	175
10.2.2 前向算法	175
10.2.3 后向算法	178
10.2.4 一些概率与期望值的计算	179
10.3 学习算法	180
10.3.1 监督学习方法	180
10.3.2 Baum-Welch 算法	181
10.3.3 Baum-Welch 模型参数估计公式	183
10.4 预测算法	184
10.4.1 近似算法	184
10.4.2 维特比算法	184
本章概要	187
继续阅读	188
习题	188
参考文献	189
第 11 章 条件随机场	191
11.1 概率无向图模型	191
11.1.1 模型定义	191
11.1.2 概率无向图模型的因子分解	193
11.2 条件随机场的定义与形式	194
11.2.1 条件随机场的定义	194
11.2.2 条件随机场的参数化形式	195
11.2.3 条件随机场的简化形式	197
11.2.4 条件随机场的矩阵形式	198
11.3 条件随机场的概率计算问题	199
11.3.1 前向-后向算法	199
11.3.2 概率计算	200
11.3.3 期望值的计算	201
11.4 条件随机场的学习算法	201

11.4.1 改进的迭代尺度法	202
11.4.2 拟牛顿法	205
11.5 条件随机场的预测算法	206
本章概要	208
继续阅读	209
习题	209
参考文献	210
第 12 章 统计学习方法总结	211
附录 A 梯度下降法	217
附录 B 牛顿法和拟牛顿法	219
附录 C 拉格朗日对偶性	225
索引	229

第1章 统计学习方法概论

本章简要叙述统计学习方法的一些基本概念，这是对全书内容的概括，也是全书内容的基础。首先叙述统计学习的定义、研究对象与方法；然后叙述监督学习，这是本书的主要内容；接着提出统计学习方法的三要素：模型、策略和算法；介绍模型选择，包括正则化、交叉验证与学习的泛化能力；介绍生成模型与判别模型；最后介绍监督学习方法的应用：分类问题、标注问题与回归问题。

1.1 统计学习

1. 统计学习的特点

统计学习 (statistical learning) 是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科。统计学习也称为统计机器学习 (statistical machine learning)。

统计学习的主要特点是：(1) 统计学习以计算机及网络为平台，是建立在计算机及网络之上的；(2) 统计学习以数据为研究对象，是数据驱动的学科；(3) 统计学习的目的是对数据进行预测与分析；(4) 统计学习以方法为中心，统计学习方法构建模型并应用模型进行预测与分析；(5) 统计学习是概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科，并且在发展中逐步形成独立的理论体系与方法论。

赫尔伯特·西蒙 (Herbert A. Simon) 曾对“学习”给出以下定义：“如果一个系统能够通过执行某个过程改进它的性能，这就是学习。”按照这一观点，统计学习就是计算机系统通过运用数据及统计方法提高系统性能的机器学习。现在，当人们提及机器学习时，往往是指统计机器学习。

2. 统计学习的对象

统计学习的对象是数据 (data)。它从数据出发，提取数据的特征，抽象出数据的模型，发现数据中的知识，又回到对数据的分析与预测中去。作为统计学习的对象，数据是多样的，包括存在于计算机及网络上的各种数字、文字、图像、视频、音频数据以及它们的组合。

统计学习关于数据的基本假设是同类数据具有一定的统计规律性，这是统计学习的前提。这里的同类数据是指具有某种共同性质的数据，例如英文文章、互联网网页、数据库中的数据等。由于它们具有统计规律性，所以可以用概率统计

方法来加以处理。比如，可以用随机变量描述数据中的特征，用概率分布描述数据的统计规律。

在统计学习过程中，以变量或变量组表示数据。数据分为由连续变量和离散变量表示的类型。本书以讨论离散变量的方法为主。另外，本书只涉及利用数据构建模型及利用模型对数据进行分析与预测，对数据的观测和收集等问题不作讨论。

3. 统计学习的目的

统计学习用于对数据进行预测与分析，特别是对未知新数据进行预测与分析。对数据的预测可以使计算机更加智能化，或者说使计算机的某些性能得到提高；对数据的分析可以让人们获取新的知识，给人们带来新的发现。

对数据的预测与分析是通过构建概率统计模型实现的。统计学习总的目标就是考虑学习什么样的模型和如何学习模型，以使模型能对数据进行准确的预测与分析，同时也要考虑尽可能地提高学习效率。

4. 统计学习的方法

统计学习的方法是基于数据构建统计模型从而对数据进行预测与分析。统计学习由监督学习 (supervised learning)、非监督学习 (unsupervised learning)、半监督学习 (semi-supervised learning) 和强化学习 (reinforcement learning) 等组成。

本书主要讨论监督学习，这种情况下统计学习的方法可以概括如下：从给定的、有限的、用于学习的训练数据 (training data) 集合出发，假设数据是独立同分布产生的；并且假设要学习的模型属于某个函数的集合，称为假设空间 (hypothesis space)；应用某个评价准则 (evaluation criterion)，从假设空间中选取一个最优的模型，使它对已知训练数据及未知测试数据 (test data) 在给定的评价准则下有最优的预测；最优模型的选取由算法实现。这样，统计学习方法包括模型的假设空间、模型选择的准则以及模型学习的算法，称其为统计学习方法的三要素，简称为模型 (model)、策略 (strategy) 和算法 (algorithm)。

实现统计学习方法的步骤如下：

- (1) 得到一个有限的训练数据集合；
- (2) 确定包含所有可能的模型的假设空间，即学习模型的集合；
- (3) 确定模型选择的准则，即学习的策略；
- (4) 实现求解最优模型的算法，即学习的算法；
- (5) 通过学习方法选择最优模型；
- (6) 利用学习的最优模型对新数据进行预测或分析。

本书以介绍统计学习方法为主，特别是监督学习方法，主要包括用于分类、标注与回归问题的方法。这些方法在自然语言处理、信息检索、文本数据挖掘等领

域中有着极其广泛的应用。

5. 统计学习的研究

统计学习研究一般包括统计学习方法 (statistical learning method)、统计学习理论 (statistical learning theory) 及统计学习应用 (application of statistical learning) 三个方面。统计学习方法的研究旨在开发新的学习方法；统计学习理论的研究在于探求统计学习方法的有效性与效率，以及统计学习的基本理论问题；统计学习应用的研究主要考虑将统计学习方法应用到实际问题中去，解决实际问题。

6. 统计学习的重要性

近 20 年来，统计学习无论是在理论还是在应用方面都得到了巨大的发展，有许多重大突破，统计学习已被成功地应用到人工智能、模式识别、数据挖掘、自然语言处理、语音识别、图像识别、信息检索和生物信息等许多计算机应用领域中，并且成为这些领域的核心技术。人们确信，统计学习将会在今后的科学发展和技术应用中发挥越来越大的作用。

统计学习学科在科学技术中的重要性主要体现在以下几个方面：

(1) 统计学习是处理海量数据的有效方法。我们处于一个信息爆炸的时代，海量数据的处理与利用是人们必然的需求。现实中的数据不但规模大，而且常常具有不确定性，统计学习往往是处理这类数据最强有力的工具。

(2) 统计学习是计算机智能化的有效手段。智能化是计算机发展的必然趋势，也是计算机技术与开发的主要目标。近几十年来，人工智能等领域的研究表明，利用统计学习模仿人类智能的方法，虽有一定的局限性，但仍然是实现这一目标的最有效手段。

(3) 统计学习是计算机科学发展的一个重要组成部分。可以认为计算机科学由三维组成：系统、计算、信息。统计学习主要属于信息这一维，并在其中起着核心作用。

1.2 监督学习

统计学习包括监督学习、非监督学习、半监督学习及强化学习。本书主要讨论监督学习问题。

监督学习 (supervised learning) 的任务是学习一个模型，使模型能够对任意给定的输入，对其相应的输出做出一个好的预测 (注意，这里的输入、输出是指某个系统的输入与输出，与学习的输入与输出不同)。计算机的基本操作就是给定一个输入产生一个输出，所以监督学习是极其重要的统计学习分支，也是统计学习中内容最丰富、应用最广泛的部分。

1.2.1 基本概念

1. 输入空间、特征空间与输出空间

在监督学习中, 将输入与输出所有可能取值的集合分别称为输入空间 (input space) 与输出空间 (output space). 输入与输出空间可以是有限元素的集合, 也可以是整个欧氏空间. 输入空间与输出空间可以是同一个空间, 也可以是不同的空间; 但通常输出空间远远小于输入空间.

每个具体的输入是一个实例 (instance), 通常由特征向量 (feature vector) 表示. 这时, 所有特征向量存在的空间称为特征空间 (feature space). 特征空间的每一维对应于一个特征. 有时假设输入空间与特征空间为相同的空间, 对它们不予区分; 有时假设输入空间与特征空间为不同的空间, 将实例从输入空间映射到特征空间. 模型实际上都是定义在特征空间上的.

在监督学习过程中, 将输入与输出看作是定义在输入 (特征) 空间与输出空间上的随机变量的取值. 输入、输出变量用大写字母表示, 习惯上输入变量写作 X , 输出变量写作 Y . 输入、输出变量所取的值用小写字母表示, 输入变量的取值写作 x , 输出变量的取值写作 y . 变量可以是标量或向量, 都用相同类型字母表示. 除特别声明外, 本书中向量均为列向量, 输入实例 x 的特征向量记作

$$x = (x^{(1)}, x^{(2)}, \dots, x^{(i)}, \dots, x^{(n)})^T$$

$x^{(i)}$ 表示 x 的第 i 个特征. 注意, $x^{(i)}$ 与 x_i 不同, 本书通常用 x_i 表示多个输入变量中的第 i 个, 即

$$x_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)})^T$$

监督学习从训练数据 (training data) 集合中学习模型, 对测试数据 (test data) 进行预测. 训练数据由输入 (或特征向量) 与输出对组成, 训练集通常表示为

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

测试数据也由相应的输入与输出对组成. 输入与输出对又称为样本 (sample) 或样本点.

输入变量 X 和输出变量 Y 有不同的类型, 可以是连续的, 也可以是离散的. 人们根据输入、输出变量的不同类型, 对预测任务给予不同的名称: 输入变量与输出变量均为连续变量的预测问题称为回归问题; 输出变量为有限个离散变量的预测问题称为分类问题; 输入变量与输出变量均为变量序列的预测问题称为标注问题.